CORPUS

Corpus

8 | 2009
Corpus de textes, textes en corpus

Marcel CORI, Sophie DAVID, Jacqueline LéON (dir.). — *Langages* n°171 « Construction des faits en linguistique : la place des corpus ». Paris : Larousse, septembre 2008, 132 pages.

Sylvie Mellet

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/1776
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 15 novembre 2009
ISSN : 1638-9808

Référence électronique
Sylvie Mellet, « Marcel CORI, Sophie DAVID, Jacqueline LÉON (dir.). — *Langages* n°171 « Construction des faits en linguistique : la place des corpus ». Paris : Larousse, septembre 2008, 132 pages. », *Corpus* [En ligne], 8 | 2009, mis en ligne le 01 juillet 2010, consulté le 19 avril 2019. URL : http://journals.openedition.org/corpus/1776

Marcel CORI, Sophie DAVID, Jacqueline LéON (dir.). — *Langages* n° 171 « Construction des faits en linguistique : la place des corpus ». Paris : Larousse, septembre 2008, 132 pages.

Sylvie Mellet

1 Cette livraison de la revue *Langages* semble avoir un objectif éditorial officiel, affiché dans son titre, consistant à porter un regard réflexif et critique sur les apports et les limites des (grands) corpus à divers domaines de l'analyse linguistique, et un objectif profond, moins immédiatement perceptible mais sans doute premier dans l'esprit des coordinateurs et particulièrement de Marcel Cori, consistant à remettre à sa juste place la linguistique de corpus accusée d'avoir des ambitions hégémoniques excessives.

2 Le premier objectif est abordé par les trois articles centraux : l'un, collectif, traite de l'usage des corpus en morphologie, plus particulièrement en morphologie constructionnelle[1] ; le deuxième, dû à Elisabeth Delais-Roussarie, s'intéresse à l'articulation entre données construites et données attestées en phonologie post-lexicale ; et le troisième, sous la plume d'Alexander Geyken, évalue l'impact de l'usage des très grands corpus sur l'élaboration des dictionnaires monolingues. On notera, en le regrettant, qu'une fois encore la syntaxe est absente de cette réflexion sur l'usage des corpus dans la construction des données linguistiques[2].

3 Fradin *et al.* énumèrent cinq arguments en faveur du recours aux corpus pour construire les données en morphologie constructionnelle : les corpus y sont nécessaires pour valider ou infirmer des hypothèses théoriques, pour contrôler les jugements d'acceptabilité (et remplacer avantageusement l'introspection, même si « acceptable » et « attesté », comme

« inacceptable » et « non attesté », ne sont bien sûr pas équivalents), pour faciliter les mesures de productivité, pour mettre en évidence des phénomènes et des régularités que seule une très grande masse de données peut donner à voir, et, à l'inverse, pour favoriser le repérage des manques et des lacunes dans un micro-système. Chacun de ces usages est ensuite illustré par les résultats d'études antérieures déjà publiées. On notera que ces cinq usages des corpus relèvent d'attitudes épistémologiques extrêmement différentes, ce que les auteurs ne soulignent pas : les deux premiers, en effet, font du corpus la pierre de touche d'une hypothèse théorique préalable – ils sont ce que la *Corpus Linguistics* appelle *corpus-based* ; les derniers, au contraire, laissent émerger du corpus des informations neuves par lesquelles le chercheur accepte de se laisser surprendre – ils sont plutôt *corpus-driven*. On peut regretter que cette opposition n'ait pas été reprise ici alors qu'elle a été bien posée, dans l'article introductif, par J. Léon, qui y voit une des sources de la division de la *Corpus Linguistics* en deux courants nettement séparés (voir plus bas). On a donc là un indice clair du grand éclectisme qui règne en général dans les pratiques françaises en la matière. Une autre trace de cet éclectisme apparaît encore dans l'article de Fradin *et al.* et concerne la définition même de ce que doit être (ou peut être) un corpus : en effet, après avoir défini en introduction les corpus au sens strict comme des collections de données construites, formant un ensemble stable, échantillonné pour être représentatif de la langue et donc pour lesquels des critères précis définissent ce qui peut y entrer et ce qui ne le peut pas, les auteurs reconnaissent bien volontiers que leur pratique courante les conduit à utiliser toutes sortes de recueils de données, sauf peut-être, justement, des corpus *stricto sensu* ! La Toile, en particulier, est pour eux une source importante de données ; or, bien évidemment, la Toile n'est pas un corpus, c'est une ressource. Il est d'ailleurs révélateur qu'au fil de l'article l'expression « données numérisées » remplace de plus en plus souvent le mot « corpus ». En revanche, l'analyse réflexive sur l'apprêt des données est extrêmement minutieuse et montre clairement à quel point chaque étape (sélection, nettoyage, étiquetage et lemmatisation) est à la fois indispensable et légitime, mais aussi risquée et source de discussions sans solution définitive. Les exemples donnés sont parlants et suggèrent l'impact de ces choix préalables sur les analyses ultérieures. Si ce paragraphe réflexif est intéressant, il n'apporte cependant pas d'éléments vraiment neufs sur la question et l'on peut regretter que l'ensemble de l'article n'approfondisse pas de manière plus pointue l'apport des grands corpus numérisés à la discipline : on comprend bien que nombre d'études morphologiques n'auraient pas été possibles sans eux, mais on ne sait pas si ces études auparavant impossibles ont fait évoluer sensiblement la théorie, ont modifié en profondeur la construction de l'objet – bref, si elles ont suscité un véritable saut épistémologique.

4 L'article d'E. Delais-Roussarie tente de réfléchir à l'articulation, en phonologie post-lexicale, entre données construites (par exemple phrases ou textes fabriqués par le linguiste et donnés à lire en contexte expérimental) et données authentiques (*i.e.* produites en situation de communication non artificielle). L'objectif est de montrer que les deux types de données sont complémentaires, que les données authentiques ne suppriment pas la nécessité d'avoir recours aux expérimentations sur données fabriquées et que, finalement, ces dernières peuvent être légitimement intégrées à un travail sur corpus entendu au sens large. Les arguments pour englober ainsi dans le concept de corpus les données construites sont principalement : (i) leur incontournable nécessité en phonologie post-lexicale ; (ii) la « représentativité » des données ainsi construites qui sont produites selon un protocole extrêmement rigoureux ; (iii) la différence entre la phrase

(ou le texte) fabriquée par le linguiste et sa réalisation phonologique par le locuteur sollicité pour le test. La nécessité des données expérimentales, notamment dans ce champ disciplinaire, n'échappe sans doute à personne ; mais cet argument de fait ne nous paraît pas justifier en soi l'extension du terme corpus proposé par l'auteure. D'autant que la démonstration de l'utilité de ces données (pp. 66-67) est typiquement une démonstration de linguiste « en chambre » (ou en cabine expérimentale) : un certain nombre de principes morphosyntaxiques et prosodiques ayant été définis préalablement (l'auteure ne précise pas comment), ils sont illustrés et simultanément validés par des exemples du type *Les enfants de Pierre sont venus* ou *Le jeune frère de Jean vient*, présentés hors tout contexte textuel ou situationnel. Le deuxième argument nous paraît confondre les notions d'échantillon représentatif et de paradigme d'énoncés (voir notamment la fin du § 2.1 p. 71) : le paradigme d'énoncés, dans lequel le linguiste peut contrôler les paramètres, les faire varier, simplifier les contraintes et les interactions entre les divers éléments en présence, est bien sûr un outil incontournable de l'analyse linguistique ; mais cette construction, ces manipulations ne permettent pas « de gagner en représentativité » (p. 71) et ne transforme pas cet ensemble d'énoncés en un échantillon : l'échantillon est un sous-ensemble qui représente en miniature et *dans les mêmes proportions* les usages effectifs de la langue. Enfin, si l'on entend bien qu'il y a une différence de statut entre la phrase écrite proposée au lecteur et sa réalisation orale par celui-ci, on doute que l'une et l'autre soient totalement indépendantes et que la seconde puisse être élément d'un corpus si la première ne peut y prétendre. En somme, cet article est très intéressant dans son analyse de la complémentarité nécessaire des données authentiques et des données construites ; en revanche, il ne convainc pas dans sa volonté d'intégrer, au nom de cette complémentarité, les données construites dans le champ définitoire du corpus[3].

5 C'est enfin l'évolution de la pratique dictionnairique qui est mise en question par A. Geyken ; celui-ci se demande en effet « ce qu'apportent les grands corpus électroniques par rapport aux « ressources traditionnelles » des lexicographes » (p. 77). L'étude est faite à partir de l'anglais et de l'allemand. Elle se présente comme très réservée sur les apports réels des grands corpus à la constitution des dictionnaires. Or si cette distance critique nous paraît de bon aloi et largement préférable à une confiance aveugle dans les vertus du progrès technologique, elle nous semble toutefois mettre en avant des arguments discutables. Le constat d'A. Geyken est que les plus grands corpus échantillonnés ne contiennent pas encore assez d'occurrences pour fournir toutes les attestations nécessaires à la rédaction d'un dictionnaire monolingue de référence et que les corpus « opportunistes », souvent plus importants parce que plus faciles à réaliser, introduisent, eux, un biais de genre ou de niveau de langue néfaste au projet lexicographique. Cela est particulièrement sensible en allemand où la composition est extrêmement productive : les corpus fourmillent donc de mots composés parfaitement transparents et sans aucun intérêt pour le lexicographe et manquent en revanche d'attestations pour certains composés plus complexes. Tout cela est vrai et n'est pas nouveau : le dilemme entre taille insuffisante du corpus équilibré et biais générique du corpus « opportuniste » a été posé depuis fort longtemps et, il est vrai, pas encore vraiment résolu. Il nous semble cependant qu'en matière de lexicographie, c'est la complémentarité des outils qui devrait être mise en avant et l'on s'étonne un peu de voir reprocher à un corpus échantillonné de ne pas fournir des attestations de mots du langage enfantin, de variantes régionales ou archaïques et de termes très techniques (p. 82). D'une part, il appartient peut-être alors au lexicographe, avant de critiquer le corpus, de s'interroger sur la pertinence d'intégrer

de telles entrées dans un dictionnaire généraliste ; d'autre part, si après réflexion il pense devoir maintenir ces entrées, il lui appartient aussi de solliciter les corpus spécialisés idoines qui lui fourniront les attestations. De la même façon, on ne peut guère s'étonner de ne pas trouver dans un corpus de presse écrite soutenue (grands quotidiens nationaux, hebdomadaires de référence) des attestations d'expressions figées familières telles que « avoir la dalle », « faire de la gratte », etc. (p. 82). Enfin, la critique concernant les variations de genre du mot emprunté *Blackout* en allemand (neutre ou masculin, dans des proportions variant extrêmement d'un corpus à l'autre) ne saurait être mise au passif des corpus (p. 86) : la variation existe, l'emprunt est mal stabilisé dans la langue d'accueil, on ne saurait en faire grief aux documents qui en attestent ! La variation dans toutes ses dimensions a toujours été un problème crucial de la lexicographie ; il vaut sans doute mieux l'affronter au travers d'attestations multiples plutôt que de s'en remettre à l'intuition ou au hasard des lectures du lexicographe. Face aux insuffisances des corpus mais aussi à leur indéniable utilité il convient sans doute de méditer encore cette phrase de Fillmore : « I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want explore ; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. »[4]

6 Revenons maintenant aux articles qui encadrent ces trois contributions et qui répondent à l'objectif exprimé en ces termes par M. Cori dans l'introduction : « Ce numéro tente d'éclairer de manière critique les fondements historiques et les enjeux épistémologiques et linguistiques sous-jacents » du développement de la linguistique de corpus (p. 8). Il s'agit, outre l'introduction assez brève, de l'article liminaire de Jacqueline Léon retraçant l'histoire de la *Corpus Linguistics* (« Aux sources de la *Corpus Linguistics* : Firth et la *London School* ») et des deux derniers articles du numéro, celui de M. Cori, « Des méthodes de traitement automatique des langues aux linguistiques fondées sur les corpus » et celui de M. Cori et S. David qui concluent en répondant à la question « Les corpus fondent-ils une nouvelle linguistique ? » (la réponse est négative).

7 L'article de J. Léon est remarquablement clair, bien documenté et utile pour situer l'ensemble de la problématique. Il offre un historique, et donc une mise en perspective, indispensables pour poser correctement le débat sur le rôle et le statut des corpus dans la linguistique contemporaine. Particulièrement intéressante est l'analyse précise des bases sur lesquelles la *Corpus Linguistics* développée par Firth et Jones s'est scindée en deux courants radicalement différents : l'un porté par Halliday, puis Sinclair, qui développent une conception probabiliste du langage et centrent leurs travaux notamment sur l'étude des *collocations* et du *meaning by collocation* ; l'autre, porté par Quirk, puis Leech, dont les objectifs plus appliqués (didactiques) les conduisent à rechercher plutôt dans les corpus des *patterns* grammaticaux, à relever les différents usages en synchronie à des fins prescriptives et à trouver les moyens d'évaluer les écarts entre les jugements d'acceptabilité des locuteurs et leurs productions réelles. Pour Sinclair, le corpus doit être une collection non finie de textes intégraux, qu'il faut augmenter sans cesse ; pour Leech, le corpus doit au contraire être échantillonné par le linguiste pour avoir la meilleure représentativité. Enfin l'un et l'autre se positionnent contre le programme chomskyen, mais de manière différente : pour Sinclair, c'est l'idée de créativité linguistique infinie qui n'est pas acceptable et incompatible avec son modèle collocationnel ; pour Leech, la critique est plus radicale encore : elle oppose le modèle probabiliste et inductif au modèle

logico-déductif, valorise la performance contre la compétence, la description et la prise en compte de la variation contre la recherche d'universaux, l'empirisme contre le rationalisme. C'est à Leech que l'on doit l'assertion selon laquelle la linguistique de corpus est une *new kind of linguistics*, assertion contre laquelle s'élève M. Cori dans les deux derniers articles.

8 L'avant-dernier article du numéro retrace en effet l'histoire du Traitement automatique des langues (TAL) pour pointer et dénoncer ses ambitions hégémoniques sur la linguistique *via* le fagocitage de la linguistique de corpus. L'auteur en veut pour preuve l'évolution terminologique de la revue *TAL*, organe de la discipline en France : « En étudiant une suite de trois numéros de la revue *TAL* parus sur huit ans (1995, 2001, 2003), on observe que des travaux dont la nature reste semblable sont présentés sous des intitulés dont la prétention s'accroît au fil du temps. Ce qui s'appelait « traitements probabilistes » devient « linguistique de corpus », puis « modélisation probabiliste du langage naturel », alors que les contenus couverts ne semblent pas réellement différents » (p. 104). Et plus loin : « La question déterminante qui se pose par conséquent est celle d'une juste caractérisation épistémologique du TAL, de la place respective du TAL et de la recherche en linguistique. Un certain nombre d'auteurs, en effet, s'emploient à ne pas distinguer clairement traitement automatique et recherche en linguistique, voire à mettre la recherche en linguistique à la remorque du traitement automatique » (p. 108). Deux choses sont gênantes dans cette critique. La première est son caractère polémique : même si l'attaque est légitime et très probablement justifiée par quelques abus de langage et quelques enjeux de pouvoir aggravés par la conjoncture actuelle de la recherche (et, sur ce point, le dernier paragraphe de la conclusion est particulièrement pertinent et bienvenu), on peut quand même s'interroger sur l'intérêt, en 2008, de repartir en guerre contre une déclaration stupide faite par un dirigeant d'IBM et rapportée dans la revue *TAL* en 1995 (p. 107). La deuxième maladresse, épistémologiquement beaucoup plus gênante, réside dans le fait que, tout en voulant dissocier le TAL de la linguistique, l'auteur valide en fait sa confusion avec la linguistique de corpus d'une part, la statistique linguistique d'autre part : c'est en effet par le biais du TAL qu'il critique l'une et l'autre. Sa posture revient donc à ignorer que des pans entiers de la recherche sur corpus, et notamment de la recherche française, revendiquent leur aptitude à exploiter les corpus et à utiliser la statistique au bénéfice d'études de linguistique générale ou d'analyse du discours, tout en refusant farouchement d'être assimilés à du TAL. Ce n'est pas parce que les linguistes de corpus et certains praticiens du TAL collaborent, partagent certains outils et que, parfois peut-être, les seconds revendiquent abusivement les avancées des premiers, qu'il faut accuser les premiers de trahison ; d'ailleurs, a-t-on jamais rejeté la linguistique formelle ou dénoncé ses prétentions à l'époque où le TAL théorique prenait appui sur elle ?

9 Malheureusement, le dernier article, co-écrit par M. Cori et S. David, continue dans la même veine. Le questionnement est certes plus large et se détache du TAL, mais l'objectif avoué de minimiser les apports de la linguistique de corpus repose trop souvent sur une caricature de celle-ci. Les auteurs lui reprochent pêle-mêle son manque d'unité, sa difficulté à définir ce qu'est véritablement un corpus, son ambition excessive, son positionnement anti-chomskyen. On avait espéré être sorti depuis quelques années de ces crispations, mais cet espoir est ici déçu. Commençons par rappeler que si la remise en cause des principes génératives est une activité condamnable, alors la linguistique de corpus ne doit pas porter seule le poids de la condamnation : depuis longtemps la

sociolinguistique d'un côté, les modèles fonctionnalistes de l'autre (pour ne citer que deux grands courants linguistiques) se sont permis d'avoir un regard critique sur un certain nombre de ces principes. Par ailleurs, ne caricaturons pas la linguistique de corpus : non, celle-ci ne considère pas « qu'un fait est une évidence brute, issue directement de l'observation » (p. 120 ; voir aussi début du § 2.2, p. 123) ! Plus que d'autres au contraire, le linguiste de corpus est confronté au problème de la construction de ses observables, précisément parce qu'il travaille sur corpus : il sait pertinemment que ses données ont été sélectionnées, rassemblées, limitées, organisées, nettoyées, etc. Et la contrainte matérielle que représentent de tels traitements a conduit nombre d'entre eux à réfléchir à ces pratiques et à produire des articles de méthodologie réflexive et souvent critique, dont la mention manque singulièrement dans la bibliographie des deux auteurs. C'est d'ailleurs ce dont atteste l'article de Fradin *et al.* dans ce même numéro de la revue. Face à ses données qui sont tout sauf brutes, le linguiste de corpus sait aussi développer des raisonnements linguistiques : ses calculs de compatibilité (p. 123) ne diffèrent de ceux des autres linguistes qu'en ce qu'ils sont obligatoirement confrontés à la diversité des paramètres contextuels qui entrent en interaction avec le phénomène étudié : la réduction simplificatrice nécessaire est donc fortement perçue et contrôlée. Il y aurait encore beaucoup à dire sur l'oubli, dans ce dernier article, du rôle positif des corpus pour le renouvellement des exemples transmis de génération en génération à travers des grammaires souvent très répétitives, ou sur la présentation simplificatrice – pour ne pas dire erronée – des objectifs de la statistique linguistique ; la discussion sur les rapports entre grammaticalité, acceptabilité et attestation devrait être reprise sur de nouvelles bases ; enfin la spécificité de la linguistique de corpus française devrait être examinée beaucoup plus précisément, à la lumière des divers héritages qu'elle assume et qui ne se réduisent pas au TAL, loin s'en faut (analyse de discours, lexicométrie, linguistique textuelle). Mais plutôt que de conclure par une polémique stérile que, précisément, nous récusons, nous dirons que ce numéro de *Langage* est une amorce stimulante pour une réflexion plus approfondie et mieux historicisée sur la mode de la ou des linguistique(s) de corpus.

NOTES

1. Article largement collectif de Bernard Fradin, Georgette Dal, Natalia Grabar, Stéphanie Lignon, Fiammetta Namer, Delphine Tribout, Pierre Zweigenbaum.

2. Signalons que le numéro 9 de la revue *Corpus*, à paraître en 2010, sera entièrement consacré à ce rapport de la syntaxe aux corpus.

3. A noter dans cet article une coquille : Meillet (2002) en page 71, au lieu de Mellet (2002), coquille d'autant plus gênante que la référence n'est pas reprise en bibliographie générale et ne peut donc pas être rectifiée.

4. In Jan Svartvik (ed.), *Directions in Corpus Linguistics*, Proceedings of Nobel Symposium 82. Berlin / New York : Mouton de Gruyter (Trends in Linguistics. Studies and Monographs 65), 1992, p. 35.

AUTEUR

SYLVIE MELLET

CNRS, Université Nice-Sophia Antipolis